



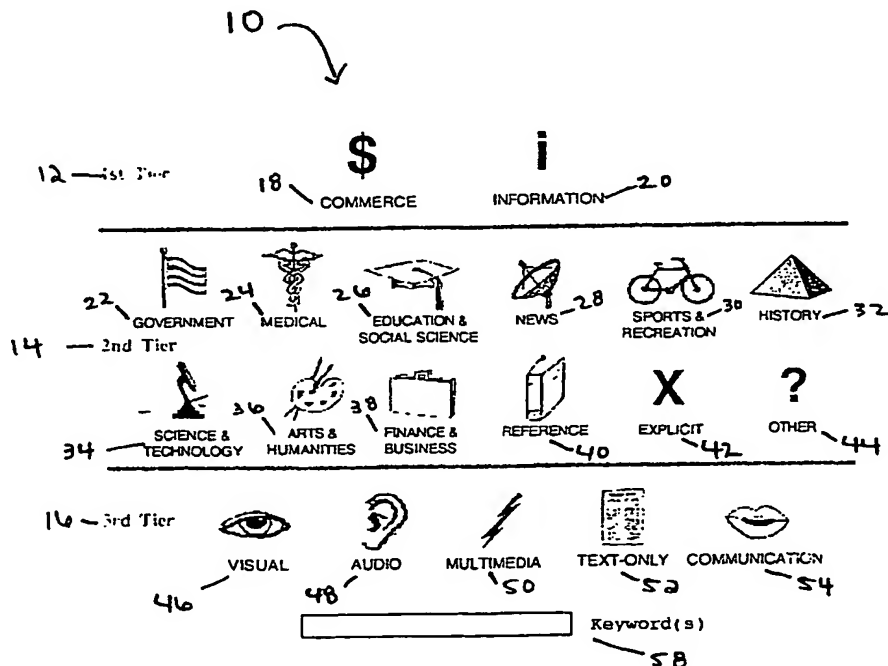
INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : G06F 17/30		A2	(11) International Publication Number: WO 00/67161
			(43) International Publication Date: 9 November 2000 (09.11.00)
(21) International Application Number: PCT/US00/12376		(81) Designated States: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).	
(22) International Filing Date: 3 May 2000 (03.05.00)			
(30) Priority Data: 60/132,694 4 May 1999 (04.05.99) US 09/565,695 3 May 2000 (03.05.00) US			
(71)(72) Applicants and Inventors: GRANT, Lee, H. [US/US]; 4849 El Cemente #169, Davis, CA 95616 (US). CAPIZZI, Susan, A. [US/US]; 4849 El Cemente #169, Davis, CA 95616 (US).			
(74) Agent: MILLEMANN, Audrey, A.; Weintraub Genshlea & Sproul, 11th floor, 400 Capitol Mall, Sacramento, CA 95814 (US).		Published Without international search report and to be republished upon receipt of that report.	

(54) Title: METHOD AND APPARATUS FOR CATEGORIZING AND RETRIEVING NETWORK PAGES AND SITES

(57) Abstract

The invention includes a method for categorizing pages on a network, including the steps of determining whether a page is involved in transacting business or providing information, has information relating to a subject matter category, and has particular types of files associated with it. The apparatus for categorizing pages on network includes mechanisms for determining whether the page is involved in transacting business or providing information, has information relating to a subject matter category, and has particular file types associated with it. The invention also includes a method for searching for information on a network. The steps include providing an opportunity to limit a search to categories including commerce and information, subject matter, and file type; and providing an opportunity to limit the search by keyword. Also included are the steps of identifying pages within the selected categories and which contain the keyword, and reporting the results of the search. The invention also includes apparatus for searching for information on a network.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Latvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakistan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

**Method and Apparatus for Categorizing and
Retrieving Network Pages and Sites**

RELATED APPLICATION

This application claims the benefit of the filing date of earlier-filed, co-pending provisional application serial no. 60/132,694 filed on May 4, 1999, pursuant to 35 U.S.C. § 119(e).

BACKGROUND OF THE INVENTION

1. Field Of The Invention

The present invention relates generally to methods and apparatus for categorizing and searching for information on a network and, more specifically, to categorizing and searching Web pages on the Internet.

2. Description Of The Related Art

The Internet contains over one billion Web pages. It has been estimated that two million Web pages are added to the Internet each day (*The Industry Standard*, February 28, 2000). This vast amount of information is a tremendous resource for the public to use. However, there is no effective way for a user to obtain relevant information. Although 85 percent of users use search engines to find information on the Internet, "a mind-boggling 92 percent of searches fail to find relevant information or to arrange the results in a meaningful order." (*The Industry Standard*, April 17, 2000, referring to a Forrester Research review of Web sites.)

There are two fundamental problems. First, there is no standardized international categorization system or catalog of the information contained on the Internet. A group of librarians and others have been working on a cataloging system for the Internet for the last few years. This work is referred to as the Dublin Core Metadata Element Set. This system suffers from a number of problems, including requiring a high degree of cataloging knowledge and being time-consuming and very expensive. In addition, because of the size of the Internet, it is a system that is unworkable.

Second, because there is no standardized categorization system or catalog, the existing search methods, which primarily include directories and search engines, are often cumbersome, ineffective, and inefficient.

Directories or indices are human-compiled databases of Web sites or pages. Most directories use editors to review and categorize Web sites. Some use contributions by their visitors. A user searches a directory by reviewing lists of categories and subcategories, or also typing in keywords. The result is a list of documents that the user can access by links. Directories are helpful to familiarize a user with the scope of a subject, but are not very useful in finding specific information. Also, directories can be slow, and the results may be haphazard.

Another major problem is that directories review and categorize only a small percentage of pages and sites. Examples of directories commonly used are Yahoo! and LookSmart.

Search engines are huge databases that automatically index large portions of the Internet and continually update that index. Search engines typically include a Web crawler or spider (also called a worm, robot, or bot) that automatically crawls through the Internet on hyperlinks indexing Web pages, a database which is the index compiled by the crawler, and a search tool which the user can use to search the database. The databases of the existing search engines differ in how they are created. Some Web crawlers index each word in a document, some index only keywords, including META tags, and some index other parts of a Web page, such as title, headings, etc. Most search engines require a search to be conducted by typing in keywords. The way in which the search query is formulated may be by Boolean logic, where keywords are used with various terms, or by natural language, where keywords are used in the form of a question. Although natural language searches may be easier for a user to formulate, both types of formulations rely on keywords.

Most search engines use mathematical algorithms to weigh or rank the results, with the most relevant items listed first. These rankings may be based on the number of times a keyword is used on a page or the location of the keyword on the page. Some search engines also allow the user to organize or group the results by category, date, or other variable, such as the folders used by Northern Light, U. S. Patent no. 5,924,090 to Krellenstein. Another search engine, known as the Clever Project, by IBM, analyzes hyperlinks between pages, in addition to text and citations, in order to develop algorithms that are intended to increase the relevancy of search results. This method is a marginal improvement over other search engines, but has its own set of problems. "A shortcoming of Clever has been that

for a narrow topic, such as Frank Lloyd Wright's house Fallingwater, the system sometimes broadens its search and retrieves information on a general subject, such as American architecture." ("Hypersearching the Web," *Scientific American*, June 1999.)

5 Search engines do not index the entire Internet. Most have indexed about one-third of the available or publicly indexable Web pages (i.e., excluding Web pages with authorization requirements). Examples of search engines are: Inktomi (the largest, with about 500 million Web pages indexed as of April 11, 2000); FAST (with about 340 million Web pages indexed); AltaVista, Northern Light, and
10 Excite. A greater portion of the Internet can be searched using a meta-search. This technology allows the user to search several search engines at the same time and presents all the results in a single list, but exacerbates the problems inherent in existing search engines.

 Because they contain such huge databases, existing search engines often
15 produce search results too voluminous for the user to review. Also, the search results typically contain a vast amount of irrelevant or unrelated items. As stated above, it has been found that 92 percent of searches did not yield relevant information or did not organize the results in a usable fashion (*The Industry Standard*, April 17, 2000). Another problem is that search engines are more likely
20 to index pages with more links, pages with commercial information, and pages in the United States, rather than lesser known, educational, or non-United States pages.

 Another major problem of existing search engines is that they may allow minors access to pornography on the Internet. Current filtering software is an
25 ineffective and often clumsy tool that fails to limit access to many pornographic sites, but blocks other sites that are educational or medical in nature. In addition, the controversy surrounding this issue has created enormous difficulties for public

libraries with respect to allowing minors access to the Internet.

Still further objects of the inventive method and apparatus disclosed herein will be apparent from the drawings and following detailed description thereof.

5

SUMMARY OF THE INVENTION

The method and apparatus for categorizing and retrieving network pages and sites of the present invention are adapted to overcome the above-noted shortcomings and to fulfill the stated needs.

10 The first embodiment of the invention is a method and apparatus for categorizing a network page. The method comprises the steps of providing a list of categories and assigning a page to one or more of a plurality of the categories. The apparatus includes means for providing a list of categories and means for assigning a page to one or more of a plurality of categories. The second embodiment of the invention is a method and apparatus for categorizing
15 pages on a network. The method comprises the steps of determining whether a page is involved in transacting business or providing information, determining whether a page has information relating to one or more of a plurality of subject matter categories, and determining the type of files associated with a page. The apparatus includes means for determining whether a page is involved in transacting
20 business or providing information, means for determining whether a page has information relating to one or more of a plurality of subject matter categories, and means for determining the type of files associated with a page. The third embodiment of the invention is a method and apparatus for searching for and locating information on a network. The method comprises the steps of providing
25 the opportunity to limit the search to categories for pages involved in transacting business, pages involved in providing information, and pages involved in both transacting business and providing information; providing an opportunity to limit

the search to one or more of a plurality of subject matter categories; providing an opportunity to limit the search to one or more of a plurality of file-type categories; and providing an opportunity to limit the search by keyword. The apparatus comprises means for providing an opportunity to limit the search to one or more of a plurality of categories, means for providing an opportunity to limit the search by keyword, means for identifying pages categorized into all the categories to which the search was limited, means for determining which of the identified pages contain the keyword to which the search was limited, and means for reporting to a user all said identified pages and keyword-containing pages.

10 It is an object of the invention to provide a method and apparatus for categorizing a page on a network, during or after the time that the page is created, according to whether the page is involved in transacting business or providing information.

15 It is an object of the invention to provide a method and apparatus for categorizing a page on a network, during or after the time that the page is created, according to the subject matter contained on the page.

It is a further object of the invention to provide a method and apparatus for categorizing a page on a network, during or after the time that the page is created, according to the type of files associated with a page.

20 It is also an object of the invention to provide a method and apparatus for searching a network containing pages that have been categorized according to whether the page is involved in transacting business or providing information, the subject matter of the page, and the type of files associated with a page.

25 It is an object of the invention to provide a method and apparatus for searching a network, such as the Internet, to allow the user access to a larger percentage of information contained on the network.

It is a further object of the invention to provide a method and apparatus for

searching a network, such as the Internet, to obtain more relevant results more quickly than existing methods for searching allow.

It is a further object of the invention to provide a method and apparatus to easily obtain audio or visual material located on a network.

5 It is another object of the invention to provide a method and apparatus for searching a network that is easy to use.

It is also an object of the invention to provide a method and apparatus that does not require the user to understand or use a particular language, including English.

10 It is a further object of the invention to provide a method and apparatus for limiting the results of a search, such as a search on the Internet, to exclude pornographic materials.

It is also an object of the invention to provide a method and apparatus with the advantages of pornography-filtering software, but without the disadvantages of
15 such software.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a representation of the preferred graphical user interface showing the three tiers and the categories within those tiers.

20 Figure 2 is a chart of the Government, Medical, News, and History categories of the second tier showing examples of topics contained within those categories.

Figure 3 is a chart of the Education & Social Sciences, Science & Technology, Sports & Recreation, and Arts & Humanities categories of the
25 second tier showing examples of topics contained within those categories.

Figure 4 is a chart of the Finance & Business, Reference, Explicit, and Other categories of the second tier showing examples of topics contained within those

categories.

Figure 5 is a Venn diagram showing the intersection of the domains corresponding to the categories of Commerce and Information.

Figure 6 is a Venn diagram showing the intersection of the domains
5 corresponding to the categories of Information and Medical.

Figure 7 is a Venn diagram showing the intersection of the domains corresponding to the categories of Information, Medical, and History.

Figure 8 is a Venn diagram showing the intersection of the domains corresponding to the categories of Information, Medical, History, and Visual.

10 Figure 9 is a diagram showing an example of the relationship between the subcategory created by selecting a combination of the categories and the keyword search.

DESCRIPTION OF THE PREFERRED EMBODIMENT

15 The invention includes methods and apparatus for categorizing a page as it is being created or as it exists on a network, and for searching a network. Networks include the Internet and private corporate networks, such as intranets and local area networks. Pages on the Internet are identifiable by unique addresses and include both Web sites and Web pages.

20 As shown on Figure 1, the invention utilizes a graphical user interface (GUI) 10, including a hierarchy of three tiers, 12, 14, and 16, to categorize, and to search for information located on, Web pages. First tier 12 is a division into one or both of two major categories: pages that are involved in transacting business and pages that are involved in providing information. In the preferred embodiment of
25 the invention, the first category 18 is designated "Commerce" and the second category 20 is designated "Information." Web pages involved in transacting business include e-commerce pages, which provide users with the ability to conduct

online purchases, sales, leases, or other financial transactions, pages that may be involved in transacting business, but do not enable the user to conduct the transaction on-line, and other pages that contain commercial information. Web pages involved in providing information include pages that contain articles, journals, publications, or other non-commercial materials. Some Web pages may be involved in both transacting business and providing information and thereby fall within both the categories of "Commerce" and "Information."

Second tier 14 is a division into one or more categories based on the subject matter the Web page contains. Many different categories can be used and many different terms may be used to identify a given category. The preferred embodiment of the invention includes twelve categories encompassing like subjects that have been carefully selected to allow users to locate and access information in an efficient manner: Government 22, Medical 24, Education & Social Science 26, News 28, Sports & Recreation 30, History 32, Science & Technology 34, Arts & Humanities 36, Finance & Business 38, Reference 40, Explicit 42, and Other 44. Each of these categories includes many topics. Figures 2, 3, and 4 list examples of the topics included in each category. For example, category 22, Government, includes the following topics: federal/state/local government, law, military, nations, politics, and taxes. Category 42, Explicit, includes pornography and sexually-explicit material. Category 44, Other, is for subjects that do not fit into any of the other categories of second tier 14.

Third tier 16 is a division into one or more categories according to the type of files associated with a Web page. There are several different types of files, including text, graphics, audio, video, multimedia, and files for communications between persons. Most search engines can recognize the type of files associated with a Web page by scanning the files and identifying the file extensions (for example, .gif, .au, .wav). The preferred embodiment of the invention includes the

following five file-type categories: Visual 46, Audio 48, Multimedia 50, Text-only 52, and Communication 54. Category 46, Visual, includes files containing pictures, charts, graphs, and diagrams. Category 48, Audio, includes files containing sound, such as music, voice, and sound effects. Category 50, Multimedia, includes files containing video, film clips, and virtual reality. Category 52, Text-only, includes files that do not contain any visual, audio, or multimedia material. Category 54, Communication, includes files containing e-mail, telnet links, ICQ, and other messaging systems.

The first embodiment of the invention is a method and apparatus for categorizing a page on a network, as the page is being created or during editing at a later time. The method includes the steps of providing the creator with a list of categories and allowing the creator to assign the page to one or more of the categories. The preferred categories are the categories of the three tiers 12, 14, and 16, as shown in Figure 1. The list of categories includes a different indicium to indicate each category. The indicium is preferably a universal symbol or icon that is not associated with any one language. The indicia preferably used are shown in Figure 1.

The creator of a Web page may assign the Web page to any number or combination of the categories of three tiers 12, 14, and 16, depending on which categories best characterize the Web page. The steps of assigning a page to categories may be performed in several different ways known to those skilled in the art. The creator may also decide not to assign the page to any of the categories of a particular tier. The outcome of the categorization method is that a page is designated to be "in" or "within" the categories that best characterize the page.

First tier 12 includes two categories: Commerce 18 and Information 20, as shown in Figure 1. The creator may assign the page to either one of the two categories of Commerce 18 or Information 20. If the page is involved in both

transacting business and providing information, the creator may assign it to both Commerce 18 and Information 20.

Second tier 14 includes twelve subject matter categories: Government 22, Medical 24, Education & Social Science 26, News 28, Sports & Recreation 30, History 32, Science & Technology 34, Arts & Humanities 36, Financial 38, Reference 40, Explicit 42, and Other 44, as shown on Figure 1. The creator may assign the page to one or more of these twelve categories.

Third tier 16 includes five file-type categories: Visual 46, Audio 48, Multimedia 50, Text-only 52, and Communication 54, as shown in Figure 1. The creator may assign the page to one or more of the five file-type categories.

After the creator decides to which categories to assign the page, the creator may mark or tag the page as belonging in or within the assigned categories by associating, with the page, the corresponding indicium for each assigned category. In addition, or alternatively, the creator may communicate the categories to which the page is assigned to one or more search engines for the purpose of allowing such search engines to locate the page, by its assigned categories, in conducting a search. The creator may change the categories during editing at a later point in time as frequently as desired.

A risk with any system whereby the creators of pages are permitted to categorize their own pages is that the creator will assign more categories to the page than are justified in order to increase the number of visitors to the page. The invention addresses this problem by including a method for verifying the accuracy of categorization of a network page. The method includes the step of scanning Web pages categorized into one or more categories, which step can be performed by a Web crawler. Pages assigned to a larger number of categories are scanned more frequently. The crawler will determine whether the page was categorized automatically, for example, by a Web crawler. If the Web page was not

categorized automatically, the Web crawler further determines whether the page was properly assigned to each such category.

The apparatus for categorizing a page includes means or mechanisms for providing a list of categories with corresponding indicia, and means for assigning
5 the page to one or more of a plurality of categories. The preferred categories are the categories of the three tiers 12, 14, and 16, as shown in Figure 1.

The second embodiment of the invention is a method and apparatus for categorizing pages on a network. This method may be performed by a Web crawler. The method includes the steps of determining whether a page is involved
10 in transacting business or providing information: assigning a business-transacting page to one category, an information-providing page to a second category, and a page that is involved in both transacting business and providing information to both the first and second categories; determining whether a page has information relating to one or more subject matter categories; assigning a page to one or more subject
15 matter categories; determining the types of files associated with a page; and assigning a page to one or more file-type categories. The method further includes the step of assigning a page that has been assigned to any two or more categories, to a subcategory that consists of only pages assigned to the identical two or more categories. The outcome of the method is that a page is determined to be "in" or
20 "within" the categories that best characterize the page.

The step of determining whether a page is involved in transacting business (i.e., is a business-transacting page) may be performed by determining whether the page includes encryption software. If the page includes encryption software, it will be determined to be involved in transacting business. Additionally, or alternatively,
25 the step may be performed by determining whether the page has the capability of permitting a user to conduct a financial transaction through the page. If so, the page will be determined to be involved in transacting business (i.e., a business-

transacting page). A page involved in providing information will be determined to be an information-providing page.

The step of assigning business-transacting pages to one category (preferably designated Commerce 18), pages involved in providing information to a second category (preferably designated Information 20), and pages that are involved in both transacting business and providing information to both categories is preferably performed by assigning business-transacting pages to a first list (containing only business-transacting pages), assigning pages involved in providing information to a second list (containing only information-providing pages), and assigning pages that are involved in both transacting business and providing information to both the first and second lists. The lists are preferably databases.

The step of determining whether a page has information relating to one or more subject matter categories is preferably performed by parsing the text of the page. There are various technologies currently available that parse text that may perform this function satisfactorily.

The step of assigning a page to one or more subject matter categories is preferably performed by assigning a page that has information related to particular subject matter categories to a separate list for each such subject matter category, where each list contains only pages having information related to that subject matter category. The categories are preferably the twelve categories of second tier 14. The lists are preferably databases.

The step of determining the type of files associated with a page may be performed by identifying files containing text, graphics, audio, video, multimedia, and communications between persons. This step can be satisfactorily accomplished by search engines that scan Web pages and recognize file extensions such as .au (audio), .wav (sound), .gif (image), .jpeg (image), .jpg (image), .avi (video), .mpg (movies), and .mpg (movies).

The step of assigning a page to one or more categories based on file type is preferably performed by assigning a page that is associated with particular file types to a separate list for each such file type, where each list contains only pages associated with a single file type. The categories are preferably the five file-type categories of third tier 16. The lists are preferably databases.

The step of assigning a page to a subcategory is performed after the page has been assigned to all possible categories from three tiers 12, 14, and 16. The Web crawler assigns a page that has been assigned to two or more categories to a subcategory consisting of only pages assigned to the identical categories. For example, a page that has been categorized into the categories of Information History, Medical, and Visual would be assigned to a subcategory containing only pages also assigned to the identical categories of Information 20, History 32, Medical 24, and Visual 46. A separate list is created for each of the possible combinations of any two or more categories of three tiers 12, 14, and 16. Each list is preferably a separate database. Examples of software that can be used for creating and managing databases are Oracle 8i version 2 with the File System option and Informix Dynamic server.

The apparatus for categorizing pages on a network includes means or mechanisms for determining whether a page is involved in transacting business or providing information; means for assigning business-transacting pages to one category, information-providing pages to a second category, and pages involved in both transacting business and providing information to both the first and second categories; means for determining whether a page has information related to one or more subject matter categories; means for assigning a page to one or more subject matter categories; means for determining the types of files associated with a page; and means for assigning the page to one or more file-type categories. The apparatus may also include means for indicating to a search engine that the page has

been categorized automatically.

The third embodiment of the invention is a method and apparatus for searching for and locating information on a network. The method allows the user to search pages on a network that have already been categorized into three tiers of categories 12, 14, and 16. The categorization may have been done by the creator of a page at the time the page was created or during editing at a later time, or by a Web crawler automatically at some time after the page was created. The method also includes a categorization step, preferably performed by a search engine, before the search is begun in order to categorize any new pages that have not yet been categorized. The categorizing step comprises assigning the page to one or more categories, including a category for pages involved in transacting business and a category for pages involved in providing information, assigning the page to one or more subject matter categories, and assigning the page to one or more file-type categories. This categorizing step may be accomplished using a Web crawler, by the method and apparatus of the second embodiment.

The method provides the user with the opportunity to limit the search by selecting categories from three tiers 12, 14, and 16 and by utilizing a keyword search. The user may select one or more categories from each of three tiers 12, 14, and 16, from one or two of the tiers, or from none of the tiers, and may or may not use the keyword search function. For convenience, as is well known in the art, when an icon is selected, its appearance changes such that it is emphasized (for example, highlighted).

The user may select, from first tier 12, the category of Commerce 18, the category of Information 20, or both categories 18 and 20. The categories may be conveniently represented on the user's screen by an icon or a symbol, for example, as is preferred: "\$" for Commerce 18 and "i" for Information 20. If the user selects "\$," the search will be restricted to only those Web pages that are

categorized as Commerce 18. This will include all pages in the Commerce category 18 as well as the subcategory that is both Commerce 18 and Information 20. Pages only in the Information category 20, and not also in Commerce 18, will automatically be excluded. If the user selects "i," the search
5 will be restricted to only those Web pages that are categorized as Information 20. This will include all pages in Information category 20 as well as the subcategory that is both Information 20 and Commerce 18. Pages only in the Commerce category 18, and not also in Information 20, will automatically be excluded. If the user selects both "\$" and "i," as shown in Figure 5, the search will be restricted to
10 only those Web pages that are categorized as both Commerce 18 and Information 20. Only subcategory 56 of Commerce and Information will be searched. Pages only in Commerce 18 and pages only in Information 20 will be excluded. If none of the categories of first tier 12 are selected, the search will include Web pages of both categories and the subcategory and will not be narrowed
15 based on whether the page is involved in transacting business or providing information.

The user next may select one or more categories from second tier 14: Government 22, Medical 24, Education & Social Science 26, News 28, Sports & Recreation 30, History 32, Science & Technology 34, Arts & Humanities 36,
20 Finance & Business 38, Reference 40, Explicit 42, and Other 44. As shown in Figure 1, each of these twelve categories may be conveniently represented on the user's screen by a different icon or symbol, for example, as is preferred: a flag for Government, a caduceus for Medical, a mortarboard for Education & Social Science, a satellite dish for News, a bicycle for Sports & Recreation, a pyramid for
25 History, a microscope for Science & Technology, an artist's palette for Arts & Humanities, a briefcase for Financial, a book for Reference, an "X" for Explicit (pornographic or sexually-explicit material), and a "?" for Other. The user may

also view a list of topics included in each category by clicking on the category. The twelve subject matter categories and their corresponding topics are shown in Figures 2, 3, and 4. If none of the categories are selected, the search will include Web pages of all twelve categories and will not be narrowed based on the subject matter contained in the page.

Next, the user may select one or more categories from third tier 16: Visual 46, Audio 48, Multimedia 50, Text-only 52, and Communication 54. As shown in Figure 1, each of the five categories may be conveniently represented on the user's screen by an icon or symbol, for example, as is preferred: an eye for Visual, an ear for Audio, a lightning bolt for Multimedia, a text page for Text-only, and a mouth for Communication. If no selection is made from this tier, the results from the search will include Web pages that are associated with file-types of text, visual, audio, multimedia, and communications and will not be narrowed based on the types of files contained on the page.

Combining categories restricts the search results to only the relevant categories and subcategories. The greater the number of categories chosen, the more refined the search and the greater the number of pages that are excluded from the search. When the user selects several categories, the user does not get results from each of those categories, but only from the subcategory that is created from the combination of the selected categories. Combining categories acts as a filtering process, eliminating irrelevant material from the search and from subsequent results. This method allows the user to exclude unwanted material, such as pornography, which is contained in Explicit category 42.

The user may next enter a keyword 58, which can be a single word or multiple words. The keyword search can be formulated by using either Boolean logic terms or natural language.

After making the selections, the user initiates the search. The symbols for

the categories selected and the keyword preferably remain visible on the user's screen during the search.

After the search is initiated, a determination is made as to whether a page is categorized. A page may have been categorized using the same categories as are available to the user to limit the search, or the site may have been categorized using different categories. The determination of whether a page is categorized is preferably performed by determining whether the page is contained or referred to on a list of categorized pages. The list may be a database or an index created automatically by a Web crawler, which contains the addresses of Web pages.

Where the network being searched contains at least one page categorized into one or more of the categories which were provided to the user to limit the search, after a user initiates a category-limited search, an identification is made of all pages that have been assigned all of the categories to which the search was limited. This may be accomplished by a search engine reviewing a database corresponding to a subcategory that is equal to the combination of categories selected by the user. If the search has been limited using keyword, an identification is made of all pages containing the keyword. If the search is both category-limited and keyword-limited, an identification is made of all pages that have been assigned to all of the categories to which the search was category-limited, which also contain the keyword.

An example of how a search works is shown in Figures 6 through 9. As shown in Figure 6, if the user selects category 20 Information from first tier 12 and category 24 Medical from second tier 14, the search and subsequent search results will be limited to subcategory 60 that is created by the combination of Information 20 and Medical 24 categories, as shown by gray area. The search results will not include pages from Information category 20 or Medical category 24 that are not contained within smaller subcategory 60.

Figure 7 shows a search in which the user selected Information 20 from

first tier 12 and History 32 and Medical 24 from second tier 14. In that case, the search and subsequent search results would be limited to subcategory 62 created by the combination of Information 20, Medical 24, and History 32 categories, as shown by the gray area. The search results will not include pages from Information 20, Medical 24, or History 32 categories that are not contained within smaller subcategory 62.

Figure 8 shows a search in which the user selected Information 20 from first tier 12, Medical 24 and History 32 from second tier 14, and Visual 46 from third tier 16. In that case, the search and subsequent search results would be limited to subcategory 64 created by the combination of Information 20, Medical 24, History 32, and Visual 46 categories, as shown by the gray area. The search results will not include pages from Information 20, Medical 24, History 32, or Visual 46 categories that are not contained within smaller subcategory 64.

Figure 9 shows a search in which the user selected Information 20 from first tier 12, Medical 24 and History 32 from second tier 14, Visual 46 from third tier 16, and the keyword 58 "Pasteur." In that case, the search and subsequent search results would be limited to the subcategory created by the combination of Information 20, Medical 24, History 32, and Visual 46 categories that contain the keyword 58 "Pasteur." The search results will not include pages from Information 20, Medical 24, History 32, and Visual 46 categories that are not contained in the subcategory.

All sites identified by the search are reported as search results to the user, by network address, such as a Web page's "uniform resource locator" (URL), so that the user can access any identified page. Other information, such as the first line, may also be reported. For each site reported, the results will show all of the symbols corresponding to all of the categories to which that page had been assigned. The results will also indicate whether the categorization step was

performed automatically (for example, by a Web crawler).

The apparatus for searching for and locating information on a network includes means or mechanisms for providing an opportunity to limit the search to one or more categories from three tiers 12, 14, and 16; means for providing an
5 opportunity to limit the search by keyword; means for identifying all pages categorized into the categories to which the search was limited which contain the keyword; and means for reporting the results to a user.

The foregoing detailed disclosure of the inventive method and apparatus is considered as only illustrative of the preferred embodiment of, and not a limitation
10 upon the scope of, the invention. Those skilled in the art will envision many other variations of the method and apparatus disclosed herein that nevertheless fall within the scope of the following claims. Alternative uses for this inventive method and apparatus may later be realized. Accordingly, the scope of the invention should be determined with reference to the appended claims and not by the examples that have
15 been given herein.

CLAIMS

1. A method of categorizing a network page, comprising the steps of:
 - a. providing a list of categories; and,
 - b. assigning a page to one or more of a plurality of said categories.
2. The method of Claim 1, wherein said categories include a category for pages involved in transacting business and a category for pages involved in providing information.
3. The method of Claim 1, wherein said categories include a plurality of categories based on subject matter.
4. The method of Claim 3, wherein said categories comprise categories related to government, medical, education and social science, news, sports and recreation, history, science and technology, arts and humanities, finance and business, reference, explicit, and other.
5. The method of Claim 1, wherein said categories include a plurality of categories based on the type of files associated with a page.
6. The method of Claim 5, wherein said categories comprise visual, audio, multimedia, text-only, and communication.
7. The method of Claim 1, wherein said categories include:
 - a. a category for pages involved in transacting business and a category for pages involved in providing information;
 - b. a plurality of categories based on subject matter; and,

- c. a plurality of categories based on the type of files associated with a page.
- 8. The method of Claim 1, further comprising the step of providing an indicium for each of said categories.
- 9. The method of Claim 8, wherein said indicium comprises an icon.
- 10. The method of Claim 1, further comprising the step of communicating said categories assigned to a page to a search engine.
- 11. A method for verifying the accuracy of categorization of a network page, comprising the steps of:
 - a. for a page that has been categorized into one or more of a plurality of categories, scanning the page; and,
 - b. determining whether said page is properly included in each of said categories.
- 12. Apparatus for categorizing a network page, comprising:
 - a. means for providing a list of categories; and,
 - b. means for assigning a page to one or more of a plurality of said categories.
- 13. A method for categorizing pages on a network, comprising the steps of:
 - a. determining whether a page is involved in transacting business or in providing information;
 - b. determining whether a page has information relating to one or more

- of a plurality of subject matter categories; and,
- c. determining the type of files associated with a page.

- 5 14. The method of Claim 13, wherein said involvement-determining step comprises the step of determining whether a page includes encryption software.
- 10 15. The method of Claim 13, wherein said involvement-determining step comprises the step of determining whether a page includes the capability of permitting a user to conduct a financial transaction through a page.
- 15 16. The method of Claim 13, further comprising the step of assigning said business-transacting pages to a first category, said information-providing pages to a second category, and pages involved in both transacting business and providing information to both said first and second categories.
- 20 17. The method of Claim 16, wherein said assigning step comprises assigning said business-transacting pages to a first list, said information-providing pages to a second list, and pages involved in both transacting business and providing information to both said first and second lists.
18. The method of Claim 16, wherein said first category consists of all pages that may be utilized in the buying, selling, or leasing of a product or service.
- 25 19. The method of Claim 13, wherein said subject matter-determining step comprises the step of parsing the text of a page.

20. The method of Claim 13, further comprising the step of assigning a page to one or more of a plurality of subject matter categories.
- 5 21. The method of Claim 20, wherein said assigning step comprises assigning a page that has information relating to a subject matter category to a list containing only pages having information relating to said subject matter category.
- 10 22. The method of Claim 20, wherein said plurality of subject matter categories comprises categories related to science and technology and medical.
23. The method of Claim 20, wherein said plurality of subject matter categories comprises categories related to finance and business and reference.
- 15 24. The method of Claim 20, wherein said plurality of subject matter categories comprises categories related to government, medical, and news.
- 20 25. The method of Claim 20, wherein said plurality of subject matter categories comprises categories related to sports and recreation and history.
26. The method of Claim 20, wherein said plurality of subject matter categories comprises categories related to education and social science, arts and humanities, and reference.
- 25 27. The method of Claim 20, wherein said plurality of subject matter categories comprises categories related to pornography or sexually-explicit material.

28. The method of Claim 20, wherein said plurality of subject matter categories comprise categories related to finance and business, government, and news.
29. The method of Claim 20, wherein said plurality of subject matter categories
5 comprises categories related to government, medical, education and social science, news, sports and recreation, history, science and technology, arts and humanities, finance and business, reference, explicit, and other.
30. The method of Claim 13, wherein said file type-determining step comprises
10 the step of identifying files containing text, graphics, audio, video, multimedia, and communications between persons.
31. The method of Claim 13, wherein said file-type determining step comprises
15 identifying file extensions comprising .au, .jpeg, .jpg, .mpg, .mpeg, .avi, .wav, and .gif.
32. The method of Claim 13, further comprising the step of assigning a page to one or more of a plurality of categories based on the file type.
- 20 33. The method of Claim 32, wherein said assigning step comprises assigning a page that contains a file type to a list containing only pages containing said file type.
34. The method of Claim 32, wherein said plurality of file-type categories
25 comprise visual, audio, multimedia, text-only, and communication.
35. A method for categorizing pages on a network, comprising the steps of:

- 5
- a. determining whether a page is involved in transacting business or involved in providing information;
 - b. assigning a business-transacting page to a first category, an information-providing page to a second category, and a page involved in both transacting business and providing information to both said first and second categories;
 - c. determining whether a page has information relating to one or more of a plurality of subject matter categories;
 - d. assigning a page to one or more of a plurality of subject matter categories;
 - e. determining the type of files contained on a page; and,
 - f. assigning a page to one or more of a plurality of categories based on the type of files associated with the page.
- 10
- 15 36. The method of Claim 35, further comprising assigning a page that has been assigned to a plurality of categories to a subcategory consisting only of pages assigned to all of said plurality of categories.
- 20 37. Apparatus for categorizing pages on a network, comprising:
- a. means for determining whether a page is involved in transacting business or in providing information;
 - b. means for determining whether a page has information relating to one or more of a plurality of subject matter categories; and,
 - c. means for determining the type of files associated with a page.
- 25
38. The apparatus of Claim 37, wherein said means for determining whether a page is involved in transacting business comprises means for determining

whether a page includes encryption software.

39. The apparatus of Claim 37, wherein said means for determining whether a page is involved in transacting business comprises means for determining whether a page includes the capability of permitting a user to conduct a financial transaction through said page.
40. The apparatus of Claim 37, further comprising means for assigning said business-transacting pages to a first category, said information-providing pages to a second category, and pages involved in both transacting business and providing information to both said first and second categories.
41. The apparatus of Claim 40, wherein said assigning means operates to assign said business-transacting pages to a first list, said information-providing pages to a second list, and pages involved in both transacting business and providing information to both said first and second lists.
42. The apparatus of Claim 40, wherein said first category consists of all pages that may be utilized in the buying, selling, or leasing of a product or service.
43. The apparatus of Claim 37, wherein said means for determining whether a page has information relating to one or more of a plurality of subject matter categories comprises means for parsing the text of a page.
44. The apparatus of Claim 37, further comprising means for assigning a page to one or more of a plurality of subject matter categories.

45. The apparatus of Claim 44, wherein said assigning means operates to assign a page that has information relating to a subject matter category to a list containing only pages having information relating to said subject matter category.
- 5
46. The method of Claim 44, wherein said plurality of subject matter categories comprises categories related to sports and recreation and history.
47. The apparatus of Claim 44, wherein said plurality of subject matter categories comprises categories related to finance and business, government, and news.
- 10
48. The apparatus of Claim 44, wherein said plurality of subject matter categories comprises categories related to education and social science, arts and humanities, and reference.
- 15
49. The apparatus of Claim 44, wherein said plurality of subject matter categories comprises categories related to science and technology and medical.
- 20
50. The apparatus of Claim 44, wherein said plurality of subject matter categories comprises categories related to finance and business and reference.
- 25
51. The method of Claim 44, wherein said plurality of subject matter categories comprises categories related to government, medical, and news.

52. The apparatus of Claim 44, wherein said plurality of subject matter categories comprises categories related to government, medical, education and social science, news, sports and recreation, history, science and technology, arts and humanities, finance and business, reference, explicit, and other.
53. The apparatus of Claim 37, wherein said means for determining the file types associated with a page comprises means for identifying files containing text, graphics, audio, video, multimedia, and communications between persons.
54. The apparatus of Claim 37, wherein said means for determining the file types contained on a page comprises means for identifying file extensions comprising .au, .jpeg, .jpg, .mpg, .mpeg, .avi, .wav, and .gif.
55. The apparatus of Claim 37, further comprising means for assigning a page to one or more of a plurality of categories based on the file type.
56. The apparatus of Claim 55, wherein said assigning means operates to assign a page that contains a file type to a list containing only pages associated with said file type.
57. The apparatus of Claim 55, wherein said plurality of file-type categories comprises visual, audio, multimedia, text-only, and communication.
58. Apparatus for categorizing a page on a network, comprising:
- a. means for categorizing a page based on whether it is involved in

transacting business or in providing information, comprising:

- i. means for determining whether a page is involved in transacting business or in providing information; and,
- 5 ii. means for assigning said business-transacting pages to a first category, said information-providing pages to a second category, and pages involved in both transacting business and providing information to both said first and second categories;
- 10 b. means for categorizing a page based on whether it has information relating to one or more of a plurality of subject matter categories, comprising:
 - i. means for determining whether a page has information relating to one or more of a plurality of subject matter categories; and,
 - 15 ii. means for assigning a page to one or more of a plurality of subject matter categories; and,
- c. means for categorizing a page based on the type of files associated with a page, comprising:
 - i. means for determining the type of files associated with a page; and,
 - 20 ii. means for assigning a page to one or more of a plurality of categories based on the type of files associated with a page.
- 59. The apparatus of Claim 58, further comprising means for assigning a page that has been assigned to a plurality of categories to a subcategory consisting
25 of only pages assigned to all of said plurality of categories.
- 60. A method for searching for and locating information on a network,

comprising the steps of:

- 5 a. providing an opportunity to limit the search to one or more of a plurality of categories, wherein the categories are pages involved in transacting business, pages involved in providing information, and pages involved in both transacting business and providing information;
 - b. providing an opportunity to limit the search to one or more of a plurality of subject matter categories;
 - 10 c. providing an opportunity to limit the search to one or more of a plurality of categories based on the type of files associated with a page; and,
 - d. providing an opportunity to limit the search by keyword.
61. The method of Claim 60, wherein the opportunity to limit the search by
15 category is exercised by a user selecting an indicium corresponding to each such category.
62. The method of Claim 60, wherein the step of providing an opportunity to
20 limit the search to one or more of a plurality of subject matter categories further comprises the step of providing a separate subject matter category for pornographic material and providing an opportunity to limit the search to categories other than said pornographic material category.
63. The method of Claim 60, wherein said subject matter categories comprise
25 categories related to government, medical, education and social science, news, sports and recreation, history, science and technology, arts and humanities, finance and business, referenced, explicit, and other.

64. The method of Claim 60, wherein said file-type categories comprise visual, audio, multimedia, text-only, and communications.
- 5 65. The method of Claim 60, before providing the opportunity to limit the search, further comprising the step of categorizing a page on a network.
- 10 66. The method of Claim 65, wherein the categorizing step comprises assigning a page to one or more of a plurality of categories, wherein the categories are pages involved in transacting business, pages involved in providing information, and pages involved in both transacting business and providing information.
- 15 67. The method of Claim 65, wherein the categorizing step comprises assigning a page to one or more of a plurality of subject matter categories.
68. The method of Claim 65, wherein the categorizing step comprises assigning a page to one or more of a plurality of categories based on the type of files associated with a page.
- 20 69. The method of Claim 65, wherein the categorizing step comprises:
- a. assigning a page to one or more of a plurality of categories, wherein the categories are pages involved in transacting business, pages involved in providing information, and pages involved in both transacting business and providing information;
 - 25 b. assigning a page to one or more of a plurality of subject matter categories; and,
 - c. assigning a page to one or more of a plurality of categories based on

the type of files associated with a page.

- 5 70. The method of Claim 69, further comprising assigning a page that has been assigned to a plurality of categories, to a subcategory consisting of only pages assigned to all of said plurality of categories.
- 10 71. The method of Claim 60, further comprising the step of, after a user initiates a category-limited search, making a determination whether a page is categorized into one or more categories to which the user has had the opportunity to limit the search.
- 15 72. The method of Claim 71, wherein the step of determining whether a page is categorized comprises the step of determining whether the page is contained on a list of categorized pages.
- 20 73. The method of Claim 60, on a network having a plurality of pages categorized into one or more categories to which the user has had an opportunity to limit the search, after a user initiates a category-limited search, further comprising the step of identifying all pages categorized into all of the categories to which the search was limited.
- 25 74. The method of Claim 73, further comprising the step of reporting to a user all said identified pages.
75. The method of Claim 60, after a user initiates a keyword-limited search, further comprising the step of identifying all pages containing the keyword to which the search was limited.

76. The method of Claim 75, further comprising the step of reporting to a user all said identified pages.
77. The method of Claim 73, wherein said category-limited search was also keyword-limited, further comprising the step of determining which of said identified pages contain the keyword to which the search was limited.
78. The method of Claim 77, further comprising the step of reporting to a user all said keyword-containing pages.
79. A method for searching for and locating information on a network, comprising the steps of:
- a. providing an opportunity to limit the search to one or more of a plurality of categories, wherein the categories are pages involved in transacting business, pages involved in providing information, and pages involved in both transacting business and providing information;
 - b. providing an opportunity to limit the search to one or more of a plurality of subject matter categories, wherein said subject matter categories are: government, medical, education and social science, news, sports and recreation, history, science and technology, arts and humanities, finance and business, reference, explicit, and other;
 - c. providing an opportunity to limit the search to one or more of a plurality of categories based on the type of files associated with a page, wherein said categories are: visual, audio, multimedia, text-only, and communications; and,
 - d. providing an opportunity to limit the search by keyword.

80. The method of Claim 79, before providing the opportunity to limit the search, further comprising the step of categorizing a page on a network.
81. The method of Claim 80, wherein the categorizing step comprises:
- 5 a. assigning a page to one or more of a plurality of categories, wherein the categories are pages involved in transacting business, pages involved in providing information, and pages involved in both transacting business and providing information;
- 10 b. assigning a page to one or more of a plurality of subject matter categories; and,
- c. assigning a page to one or more of a plurality of categories based on the type of files associated with a page.
82. The method of Claim 81, further comprising assigning a page that has been
- 15 assigned to a plurality of categories, to a subcategory consisting of only pages assigned to all of said plurality of categories.
83. The method of Claim 79, on a network having a plurality of pages categorized into one or more categories to which the user has had an
- 20 opportunity to limit the search, after a user initiates a category-limited search, further comprising the step of identifying all pages categorized into all of the categories to which the search was limited.
84. The method of Claim 83, further comprising the step of reporting to a user
- 25 all said identified pages.
85. The method Claim 79, after a user initiates a keyword-limited search, further

comprising the step of identifying all pages containing the keyword to which the search was limited.

5 86. The method Claim 85, further comprising the step of reporting to a user all said identified pages.

10 87. The method of Claim 83, wherein said category-limited search was also keyword-limited, further comprising the step of determining which of said identified pages contain the keyword to which the search was limited.

88. The method of Claim 87, further comprising the step of reporting to a user all said keyword-containing pages.

15 89. Apparatus for searching for and locating information on a network, comprising:

- 20 25
- a. means for providing an opportunity to limit the search to one or more of a plurality of categories;
 - b. means for providing an opportunity to limit the search by keyword;
 - c. means for identifying all pages categorized into all of the categories to which the search was limited;
 - d. means for determining which of said identified pages contain said keyword to which the search was limited;
 - e. means for reporting to a user all said identified pages; and,
 - f. means for reporting to a user all said keyword-containing identified pages.

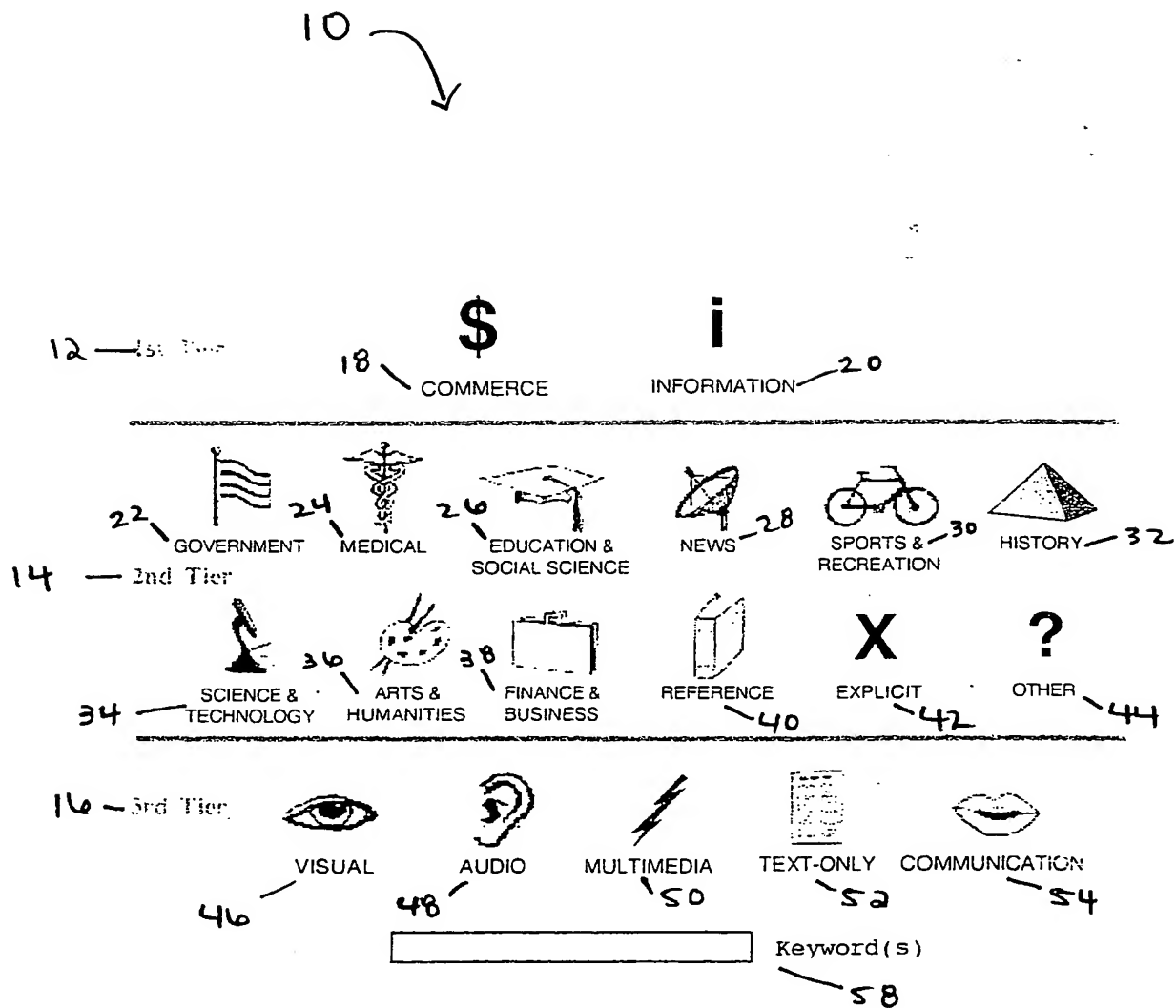


Figure 1



GOVERNMENT:

Federal/state/local government
Law
Military
Nations
Politics
Taxes



NEWS:

Internet news
News magazines
Newspapers
Radio news
Satellite news
Television news



MEDICINE:

Medicine & treatments
Alternative medicine &
treatments
Dentistry
Disabilities
Diseases
Health
Home care
Hospices
Hospitals
Medical Equipment
Nursing/Nursing homes
Nutrition
Pharmaceuticals



HISTORY:

History of people, places and
things

Figure 2



EDUCATION & SOCIAL SCIENCES:

Adult education
 Anthropology
 Archeology
 Colleges & universities
 Dissertations
 Economics
 Folklore
 Genealogy
 Languages
 Philosophy
 Political Science
 Psychology
 Public & private schools
 Religion
 Research
 Social issues
 Statistics



SCIENCE & TECHNOLOGY:

Aeronautics & space
 Agriculture
 Astronomy
 Biology
 Botany
 Chemistry
 Computers & Internet
 Ecology
 Electronics
 Engineering
 Geography (maps and atlases)
 Mathematics
 Oceanography
 Paleontology & paleozoology
 Physics
 Technical processes
 Technological advances
 Weather
 Zoology



SPORTS & RECREATION:

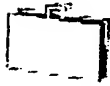
Camping
 Crafts
 Food & cooking
 Games & toys
 Gardening
 Hobbies
 Home improvements
 Pets
 Sports (football, soccer,
 baseball, etc.)
 Theme parks
 Travel



ARTS & HUMANITIES:

Architecture
 Dance
 Fine arts (painting,
 sculpture, etc.,)
 Landscape architecture
 Literature
 Movies
 Music
 Performing arts
 Photography
 Television
 Theater
 Video

Figure 3



FINANCE & BUSINESS:

Banks & financial institutions
Business information
Companies
Consumers
Employment
Finance
Investments
Marketing
Real estate
Trade
Transportation
World currencies

X

EXPLICIT:

Pornography
Sexually explicit material



REFERENCE:

Almanacs
Associations
Conferences
Dictionaries
Encyclopedias
Journals
Libraries
Museums
Organizations
Phone numbers & addresses
Proceedings
Quotations

?

OTHER:

Figure 4

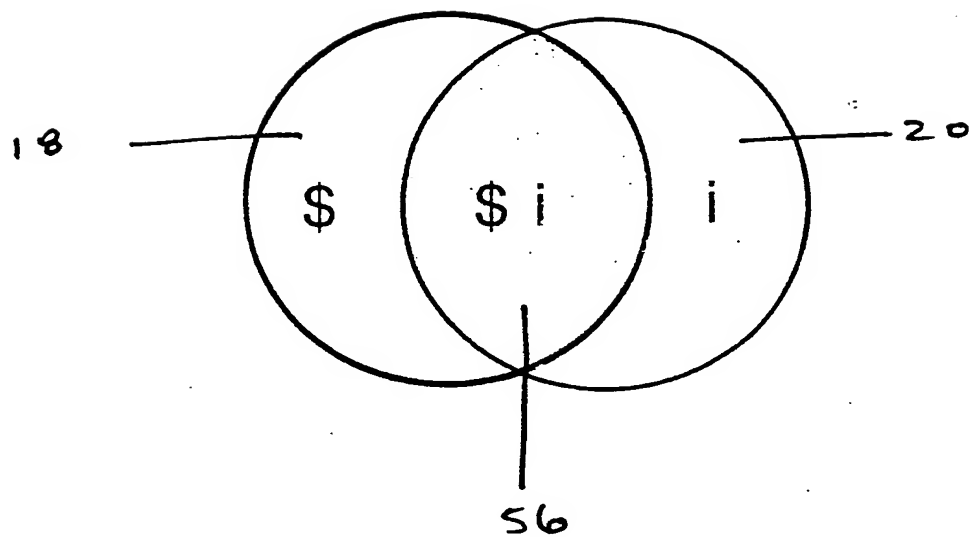


Figure 5

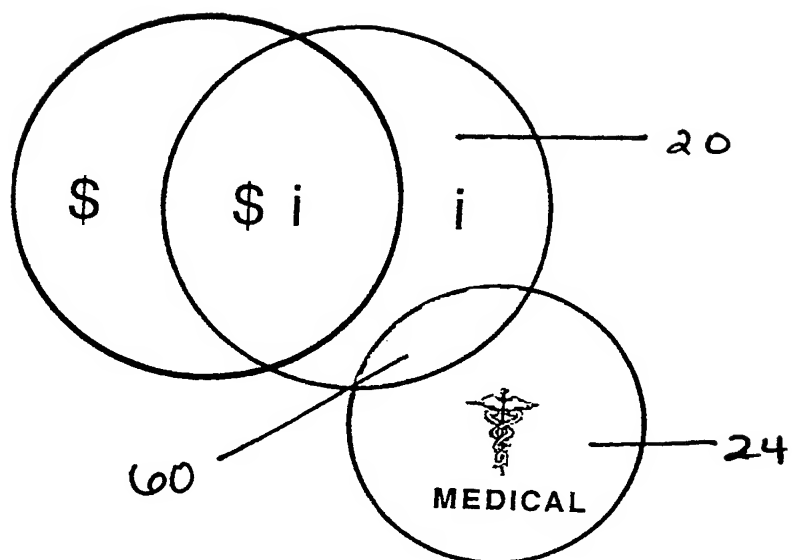


Figure 6

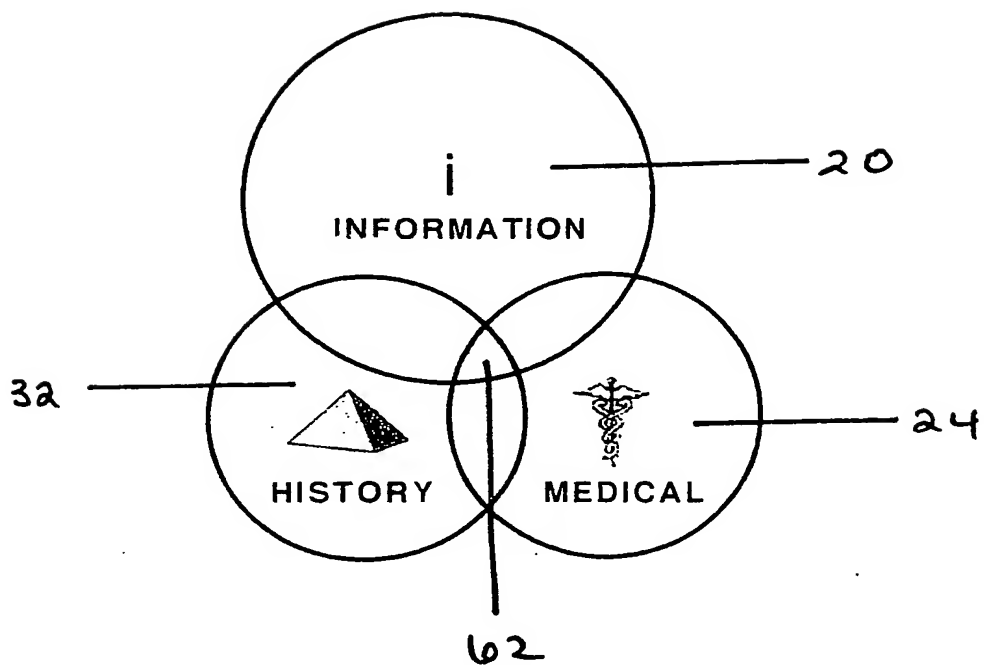


Figure 7

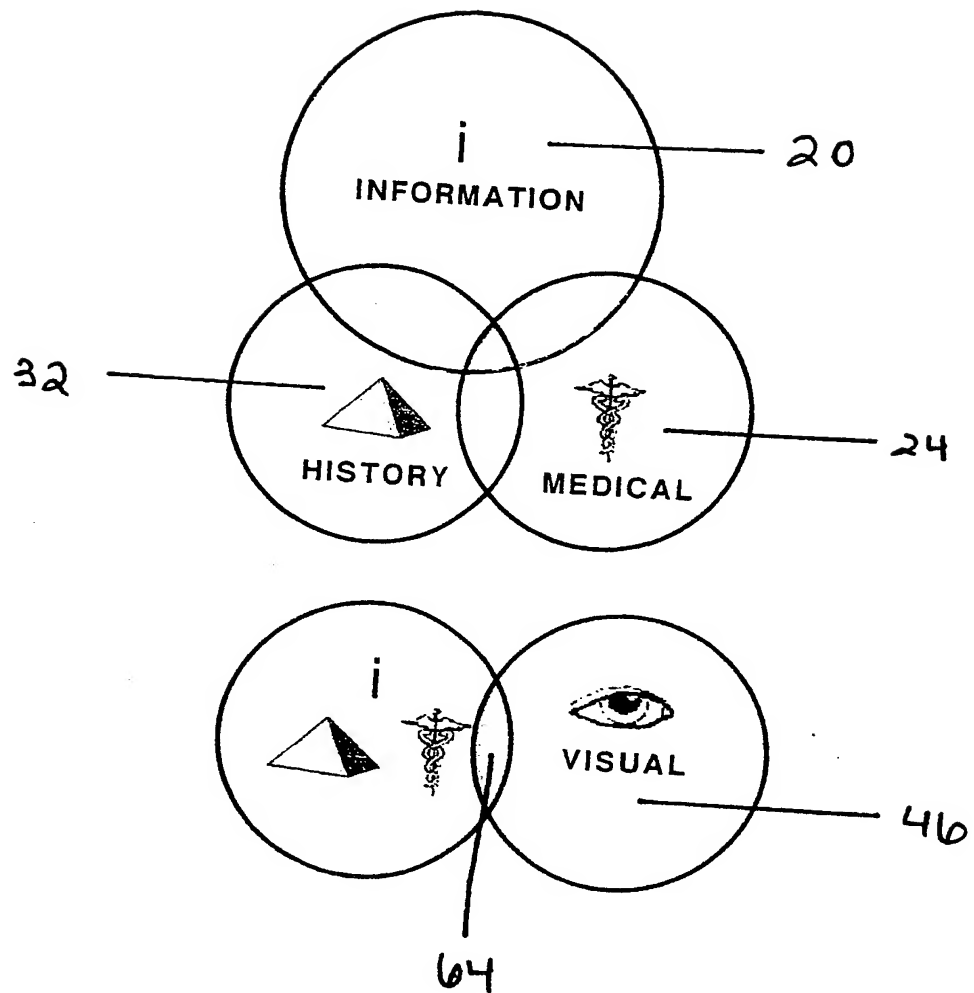
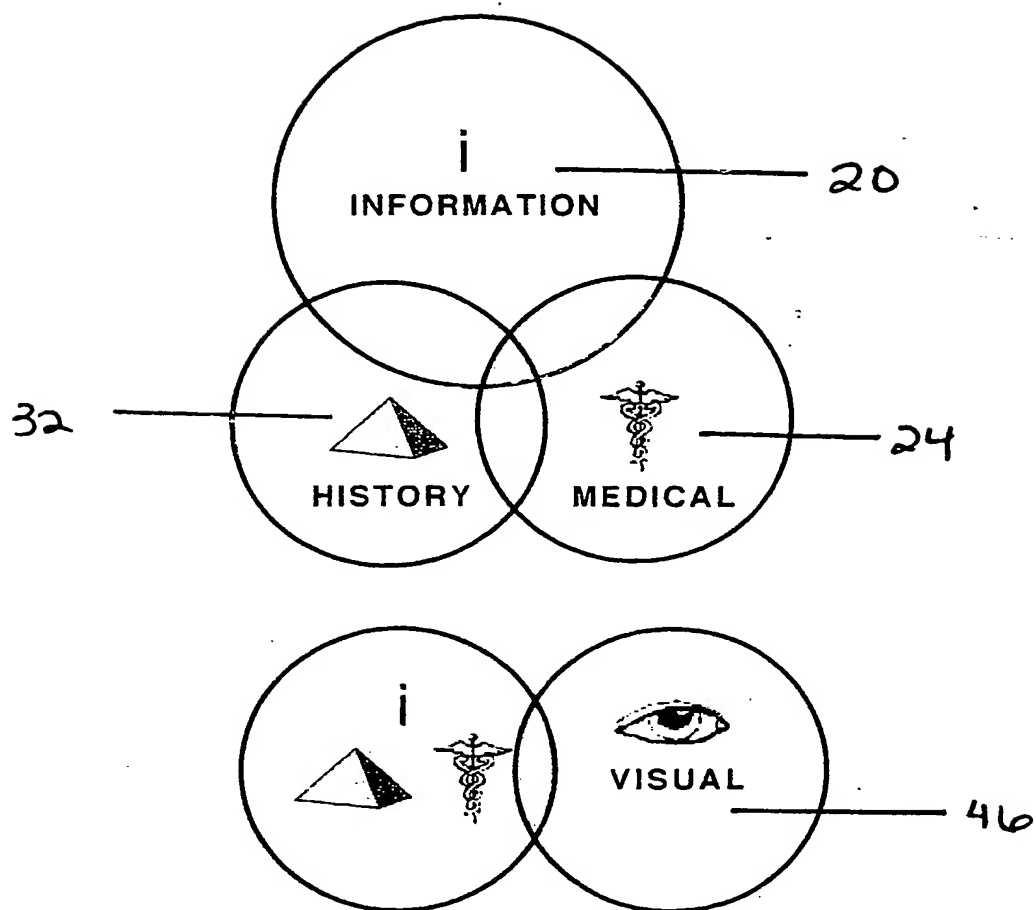


Figure 8



PASTEUR

KEYWORD(S)

58

Figure 9

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
9 November 2000 (09.11.2000)

PCT

(10) International Publication Number
WO 00/67161 A3

(51) International Patent Classification⁷: G06F 17/30

(21) International Application Number: PCT/US00/12376

(22) International Filing Date: 3 May 2000 (03.05.2000)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/132,694 4 May 1999 (04.05.1999) US
09/565,695 3 May 2000 (03.05.2000) US

(71) Applicants and

(72) Inventors: GRANT, Lee, H. [US/US]: 4849 El Cemente
#169, Davis, CA 95616 (US). CAPIZZI, Susan, A.
[US/US]: 4849 El Cemente #169, Davis, CA 95616 (US).

(74) Agent: MILLEMANN, Audrey, A.: Weintraub Genshlea
& Sproul, 11th floor, 400 Capitol Mall, Sacramento, CA
95814 (US).

(81) Designated States (*national*): AE, AG, AL, AM, AT, AU,
AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE,
DK, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID,
IL, IN, IS, JP, KE, KG, KP, LR, LS, LT, LU, LV, MA, MD,
MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD,
SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ,
VN, YU, ZA, ZW.

(84) Designated States (*regional*): ARIPO patent (GH, GM,
KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent
(AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent
(AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU,
MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM,
GA, GN, GW, ML, MR, NE, SN, TD, TG).

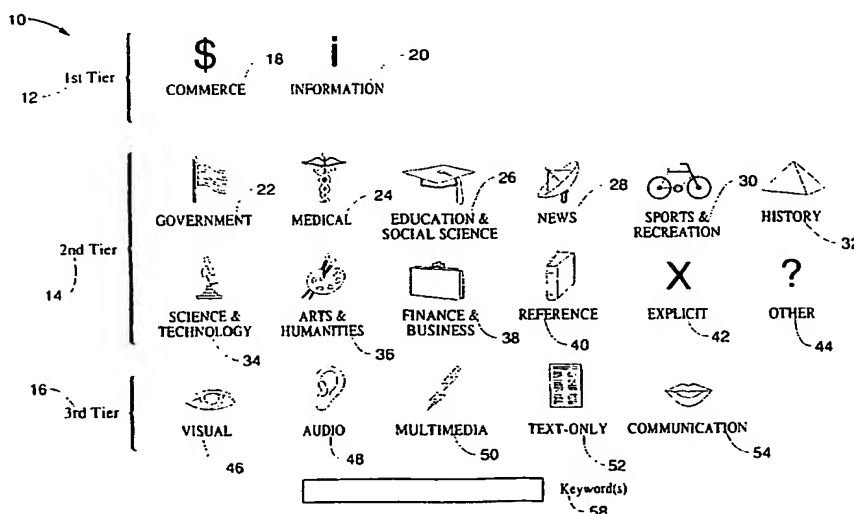
Published:

— with international search report

(88) Date of publication of the international search report:
6 June 2002

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD AND APPARATUS FOR CATEGORIZING AND RETRIEVING NETWORK PAGES AND SITES



(57) Abstract: The invention includes a method for categorizing pages on a network, including the steps of determining whether a page is involved in transacting business or providing information, has information relating to a subject matter category, and has particular types of files associated with it. The invention also includes a method for searching for information on a network. The steps include providing an opportunity to limit a search to categories including commerce and information, subject matter, and file type; and providing an opportunity to limit the search by keyword. Also included are the steps of identifying pages within the selected categories and which contain the keyword, and reporting the results of the search. The invention also includes apparatus for searching for information on a network.

WO 00/67161 A3

INTERNATIONAL SEARCH REPORT

Intern. Application No
PCT/US 00/12376

A. CLASSIFICATION OF SUBJECT MATTER
IPC 7 G06F17/30

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC 7 G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, PAJ, IBM-TDB, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CHEN H ET AL: "INTERNET CATEGORIZATION AND SEARCH: A SELF-ORGANIZING APPROACH" JOURNAL OF VISUAL COMMUNICATION AND IMAGE REPRESENTATION, ACADEMIC PRESS, INC, US, vol. 7, no. 1, 1 March 1996 (1996-03-01), pages 88-102, XP000619822 ISSN: 1047-3203	1-4, 12, 13, 19-29, 89
A	page 90, right-hand column, line 7 -page 91, right-hand column, line 33	11, 35, 37, 58, 79
X	ANONYMOUS: "Taxonomized Web Search" IBM TECHNICAL DISCLOSURE BULLETIN, IBM CORP. NEW YORK, US, vol. 40, no. 5, 1 May 1997 (1997-05-01), pages 195-196, XP002133594 ISSN: 0018-8689	1-3, 7, 12, 13, 15-29
A	the whole document	11, 35, 37, 58, 79

-/--

☒ Further documents are listed in the continuation of box C.

☒ Patent family members are listed in annex.

* Special categories of cited documents:

- *A* document defining the general state of the art which is not considered to be of particular relevance
- *E* earlier document but published on or after the international filing date
- *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- *O* document referring to an oral disclosure, use, exhibition or other means
- *P* document published prior to the international filing date but later than the priority date claimed

- *T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
- *X* document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
- *Y* document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.
- *8* document member of the same patent family

Date of the actual completion of the international search

26 October 2001

Date of mailing of the international search report

05/11/2001

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl.
Fax: (+31-70) 340-3016

Authorized officer

Deane, E

INTERNATIONAL SEARCH REPORT

Intern. Patent Application No

PCT/US 00/12376

C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT

Category *	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	<p>"SIMPLIFICATION OF A DATABASE QUERY THROUGH THE USE OF A CATEGORY WINDOW"</p> <p>IBM TECHNICAL DISCLOSURE BULLETIN, IBM CORP. NEW YORK, US,</p> <p>vol. 33, no. 38,</p> <p>1 August 1990 (1990-08-01), pages 459-461, XP000124421</p> <p>ISSN: 0018-8689</p> <p>the whole document</p>	60
A	<p>BEERUD SHETH ET AL: "EVOLVING AGENTS FOR PERSONALIZED INFORMATION FILTERING"</p> <p>PROCEEDINGS OF THE CONFERENCE ON ARTIFICIAL INTELLIGENCE FOR APPLICATIONS. ORLANDO, MAR. 1 - 5, 1993, LOS ALAMITOS, IEEE COMP. SOC. PRESS, US,</p> <p>vol. CONF. 9, 1 March 1993 (1993-03-01), pages 345-352, XP000379626</p> <p>Florida, USA</p> <p>ISBN: 0-8186-3840-0</p> <p>abstract</p> <p>page 347; right-hand column, line 1 - line 21</p>	11
A	<p>US 5 706 507 A (SCHLOSS ROBERT JEFFREY)</p> <p>6 January 1998 (1998-01-06)</p> <p>abstract</p> <p>column 2, line 41 -column 3, line 7</p> <p>column 4, line 57 -column 6, line 6;</p> <p>figures 2,3</p>	2,4,13, 35,37, 58,79
P,X	<p>US 5 924 090 A (KRELLENSTEIN MARC F)</p> <p>13 July 1999 (1999-07-13)</p> <p>cited in the application</p> <p>& WO 98 49637 A</p> <p>5 November 1998 (1998-11-05)</p> <p>page 3, line 7 - line 13</p>	1,12,89

INTERNATIONAL SEARCH REPORT

Information on patent family members

International Application No

PCT/US 00/12376

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
US 5706507	A	06-01-1998	NONE	
US 5924090	A	13-07-1999	AU 736428 B2	26-07-2001
			AU 7271798 A	24-11-1998
			EP 0979470 A1	16-02-2000
			WO 9849637 A1	05-11-1998

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

This Page Blank (uspto)